Making confidential data part of reproducible research

Lars Vilhuber (Cornell University)

WORKSHOP ON TRANSPARENCY AND REPRODUCIBILITY IN FEDERAL STATISTICS
Committee on National Statistics

Washington, 21 June 2017
Acknowledgements

Based on many discussions with numerous people

• Co-author Carl Lagoze (University of Michigan)
• Ian Schmutte (U Georgia)
• John Abowd (Cornell University, Census Bureau)
• Shawn Klimek, Lucia Foster, Danielle Foster, John Eltinge, Simson Garfinkel, many others (Census Bureau)
• Maggie Levenstein (U Michigan), Jerry Reiter (Duke)
• Many others
• … but all opinions expressed here are mine
In social and statistical sciences

- **Replicability** of research is an increasingly required part of research output.
- Replicability is often perceived as **problematic** [papers, AEA sessions, etc.]
- Replicability of **proprietary** (read: restricted access) data is seen as problematic at best, impossible at worst.
Characteristics of a good replication archive

• Permanent URL
• Identification and Availability of
  – Original data (ideally with provenance)
  – Transformed data (and the programs to transform it)
  – Broad availability
• Availability of analysis programs
By current practice

- Replication archives are created by depositing all of the materials at journals
  - Original data (optional)
  - Transformed data (unless waiver given)
  - Programs (in all cases)
Figure 1: A Breakdown of the Articles

Total Articles (109)

Confidential Data (44)  Non Confidential Data (65)

Unsuccessful (28)  Successful (37)
It is argued that...

• In order to be replicable, the data used by researchers has to be deposited at journals and/or archives

• Journals cannot handle restricted-access data, so they grant waivers = nothing happens…

• Archives can handle restricted-access deposits
Available archives

• Social and behavioral sciences
Options are available

- “Research data”

Open source research data repository software

- **Researchers**
  - Enjoy full control over your data. Receive web visibility, academic credit, and increased citation counts. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

- **Journals**
  - Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an unbreakable link between articles in your journal and associated data. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

- **Developers**
  - Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. Integrate research analysis, visualization, and exploration tools, or other research and data archival systems with Dataverse. Want to contribute?

- **Institutions**
  - Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community’s data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?
But many “proprietary datasets” cannot be deposited

- Legal restrictions (or interpretations thereof)
- The agreement is between one person and the firm, non-transferable
- The lawyers got cold feet
- Data access costs money
- The school system has no funds to archive the data, and the data were destroyed
Reproducibility with confidential data is hard
But
FEDERAL confidential data can!

(more generally: data in National Statistical Offices)
The key

- Archives are **well managed** (data persists)
- Data access to federal statistical data is **well documented** (legally and technically)
- Release of research results is **well documented** (in particular for FSRDC)
Characteristics of a good replication archive

- **Permanent URL**
- **Identification and Availability of**
  - Original data
  - Transformed data
  - Broad availability
- **Availability of analysis programs**

Project proposal → Original data → Analysis programs → Published results

All programs are in well-defined location
Characteristics of a good replication archive

- **Permanent URL**
- **Identification and Availability of**
  - Original data
  - Transformed data
  - Broad availability
- **Availability of analysis programs**

This is currently a problem

Access to data is non-discriminatory
This data is already being collected

Disclosure review request (DRR)

- Every research project with publications has at least one DRR
- The DRR captures the information necessary for a replication
In theory
What are the issues that need to be addressed?

• **Access**
  – Long lead times – dissuasive for replications
  – Capacity constraint (size, distance, computing)

• **Transparency**
  – While collected, the replication information is not easily available (absence of permanent URL)

• **Documentation**
Some suggestions

- **Access**: find a consensus among data custodians that replication is a legitimate justification for access (speed up access for replications)

\[0 \rightarrow +3 \text{ mth} \rightarrow +3 \text{ mth} \rightarrow +3 \text{ mth} = 9 \text{ mths}\]

Proposal development → Proposal review → Security clearance
Some suggestions

• **Access**: find a consensus among data custodians that replication is a legitimate justification for access (speed up access for replications)

• **Transparency**: provide assistance to users on how to cite data, describe processing, prepare replication packages

• **Leadership**: encourage and support replication activities, replication as a part of standard operating procedures
Thank you!
Questions?

lars.vilhuber@cornell.edu
Extra slides
Some are quite commendable
**Even detailed information**

<table>
<thead>
<tr>
<th><strong>Full Title</strong></th>
<th>General Social Survey, 2011: Cycle 25, Family</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subtitle</strong></td>
<td>Cycle 25, Family</td>
</tr>
<tr>
<td><strong>Alternative Title</strong></td>
<td>GSS 2011: Family</td>
</tr>
<tr>
<td><strong>Parallel Title</strong></td>
<td>Enquête sociale générale, 2011: Cycle 25, Famille</td>
</tr>
<tr>
<td><strong>Identification Number</strong></td>
<td>ca-statcan-58196</td>
</tr>
<tr>
<td><strong>Authoring Entity</strong></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Statistics Canada StatCan</td>
</tr>
<tr>
<td><strong>Producer</strong></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Statistics Canada StatCan</td>
</tr>
<tr>
<td><strong>Copyright</strong></td>
<td>Copyright © Statistics Canada, 2012</td>
</tr>
<tr>
<td><strong>Date of Distribution</strong></td>
<td>2012-07-18</td>
</tr>
<tr>
<td><strong>Series Information</strong></td>
<td>General Social Survey - Family (GSS) [4501]</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>15769.6</td>
</tr>
</tbody>
</table>
Combining Knowledge: Citations

Lars Vilhuber

ORCID ID
id: orcid.org/0000-0001-5733-8932

- Education (3)
- Employment (1)
- Funding (7)
- Works (29)

CED²AR: The Comprehensive Extensible Data Documentation and Access Repository
IEEE/ACM Joint Conference on Digital Libraries
2014-09 | conference-paper
DOI: 10.1109/jcdl.2014.6970178
Source: CrossRef Metadata Search
Preferred source